Express Mail Label No. EV402808290US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANT:	TAEJOON KWON	)
		)
FOR:	SYSTEM AND METHOD FOR DESIGNING PROBES USING HETEROGENEOUS GENETIC INFORMATION, AND COMPUTER READABLE MEDIUM	))))

CLAIM FOR PRIORITY

Mail Stop Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

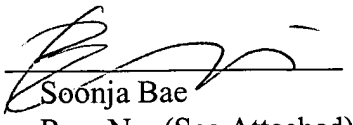
Dear Commissioner:

Enclosed herewith is a certified copy of Korean Patent Application No. 2003-0007122 filed on February 5, 2003. The enclosed Application is directed to the invention disclosed and claimed in the above-identified application.

Applicant hereby claims the benefit of the filing date of February 5, 2003, of the Korean Patent Application No. 2003-0007122, under provisions of 35 U.S.C. 119 and the International Convention for the protection of Industrial Property.

Respectfully submitted,

CANTOR COLBURN LLP

By: 

Soonja Bae
Reg. No. (See Attached)
Cantor Colburn LLP
55 Griffin Road South
Bloomfield, CT 06002
PTO Customer No. 23413
Telephone: (860) 286-2929
Fax: (860) 286-0115

Date: February 5, 2004

대 한 민 국 특 허 청
KOREAN INTELLECTUAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Intellectual Property Office.

출 원 번 호 : 10-2003-0007122
Application Number

출 원 년 월 일 : 2003년 02월 05일
Date of Application FEB 05, 2003

출 원 인 : 삼성전자주식회사
Applicant(s) SAMSUNG ELECTRONICS CO., LTD.

2003 년 03 월 07 일

특 허 청
COMMISSIONER

【서지사항】

【서류명】	특허출원서
【권리구분】	특허
【수신처】	특허청장
【참조번호】	0007
【제출일자】	2003.02.05
【국제특허분류】	G06F
【발명의 명칭】	이형 유전정보를 이용한 프로브 어레이 설계 시스템 및 방법
【발명의 영문명칭】	System for designing probe array using heterogeneneous genomic information and method of the same
【출원인】	
【명칭】	삼성전자 주식회사
【출원인코드】	1-1998-104271-3
【대리인】	
【성명】	이영필
【대리인코드】	9-1998-000334-6
【포괄위임등록번호】	2003-003435-0
【대리인】	
【성명】	이해영
【대리인코드】	9-1999-000227-4
【포괄위임등록번호】	2003-003436-7
【발명자】	
【성명의 국문표기】	권태준
【성명의 영문표기】	KWON,Tae Joon
【주민등록번호】	750512-1000311
【우편번호】	135-110
【주소】	서울특별시 강남구 압구정동 현대아파트 92동 205호
【국적】	KR
【심사청구】	청구
【취지】	특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 출원심사 를 청구합니다. 대리인 이영필 (인) 대리인 이해영 (인)

【수수료】

【기본출원료】	20	면	29,000 원
【가산출원료】	8	면	8,000 원
【우선권주장료】	0	건	0 원
【심사청구료】	12	항	493,000 원
【합계】	530,000 원		
【첨부서류】	1. 요약서·명세서(도면)_1통		

【요약서】

【요약】

이형 유전정보를 이용한 프로브 어레이 설계 시스템 및 방법이 개시된다. 저장부에는 설계 시스템 외부에 존재하는 게놈서열의 버전별 갱신이력이 기록된 크로스링크 맵이 저장된다. 정보검색부는 크로스링크 맵으로부터 게놈서열을 구성하는 유전정보중에서 대상유전정보의 식별자와 관련있는 유전정보의 식별자 및 서열정보를 획득한다. 위치추정부는 개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하고 크로스링크 맵을 기초로 기준유전정보의 시작 및 종료위치의 차이값를 계산하여 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 대상유전정보의 위치로 결정한다. 본 발명에 따르면, 최근에 설계된 프로브에 대해서 최신의 정보를 제공할 수 있으며, 프로브를 설계할 당시의 게놈서열정보 및 식별자정보가 현재 시점에서 최신의 정보가 아니더라도 크로스링크 맵을 이용하여 최신의 정보를 찾아낼 수 있다.

【대표도】

도 1

【색인어】

크로스링크 맵, 게놈, 유전정보, 프로브, 염기서열

【명세서】

【발명의 명칭】

이형 유전정보를 이용한 프로브 어레이 설계 시스템 및 방법{System for designing probe array using heterogeneneous genomic information and method of the same}

【도면의 간단한 설명】

도 1은 본 발명에 따른 프로브 어레이 설계 시스템에 대한 일 실시예의 구성을 도시한 블록도,

도 2는 크로스링크 맵의 일 예를 도시한 도면,

도 3은 위치추정부의 상세한 구성을 도시한 블록도,

도 4는 본 발명에 따른 프로브 설계 방법의 수행과정을 도시한 흐름도,

도 5는 본 발명에 따른 프로브 설계 방법의 위치추정과정을 도시한 흐름도,

그리고,

도 6은 대상유전정보가 BRCA2일 때, 크로스링크 맵에서 얻어진 이전 서열과 최신 서열의 식별자정보의 시작 및 종료위치를 도시한 도면이다.

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<7> 본 발명은 프로브 어레이 설계 시스템 및 방법에 관한 것으로, 보다 상세하게는 이형 데이터 소스를 이용하여 올리고뉴클레이티드 프로브 어레이를 설계하는 시스템 및 방법에 관한 것이다.

<8> 올리고뉴클레오티드(Oligonucleotide)를 이용한 마이크로어레이(microarray)는 적은 실험으로 많은 생물학적 정보를 얻을 수 있다는 장점으로 인하여 최근 많은 관심을 받고 있는 기술이다. 이는 기판(support material)에 프로브로서 작용하는 올리고뉴클레오티드를 부착시키고, 생물학적 정보를 얻고자 하는 샘플을 증폭 또는 레벨링 등을 통해 가공하여 얻어진 유전자가 기판에 부착된 올리고뉴클레오티드와 이루는 결합 정도를 판단하여 정보를 얻어내는 방법이다. 현재 널리 사용되는 응용 분야는 유전자들의 발현 양상을 살펴보는 분야(Expression profile)와 게놈(genome) 상의 특정 유전 정보를 확인하는 분야(genotyping)로 나누어 생각해 볼 수 있다. 두 경우 모두 샘플의 유전자에 비해서 길이가 짧은 올리고뉴클레오티드를 프로브로 사용하고, 프로브의 개수가 많다는 점에 있어서는 공통점을 가진다.

<9> 올리고뉴클레오티드를 이용하는 마이크로어레이를 만들기 위한 핵심 기술 가운데 하나는 효과적인 올리고뉴클레오티드 프로브를 선정하는 기술이다. 일반적으로, 샘플의 유전자와 샘플의 유전자에 결합할 것으로 예상되는 프로브 사이의 결합강도를 예측하여 목적에 맞는 프로브를 선별하는 방법이 사용된다. 정확한 프로브의 성능예측에 의해 적은 노력으로 성능이 양호한 마이크로어레이를 제작할 수 있으므로, 보다 정확하게 결합강도를 예측할 수 있는 방법들과 이를 바탕으로 좋은 프로브들을 선별하는 방법에 대해서 많은 연구가 진행되고 있다.

<10> 하지만 선별된 프로브들의 관련 정보를 관리하는 문제는 그리 간단하지 않다. 특히, 인간의 게놈 서열은 아직 작성중에 있기 때문에 계속 수정되고 있다. 따

라서, 이와 관련된 유전자 정보(동질이상과 같은 변이정보 및 기능에 관계된 정보 등)는 앞으로 지속적으로 변경될 것으로 예상되고 있다. 이런 상황에서 한번 설계된 프로브는 시간이 지남에 따라 설계된 당시와는 다른 정보에 관계될 수 있다. 예를 들면, 새로운 동질이상(polymorphism)이 밝혀졌다든지, 새롭게 서열이 알려진 부분이 프로브와 결합할 가능성이 있다든지 하는 일이 발생할 수 있다. 이런 경우 프로브의 설계 당시의 정보와 최신의 정보 사이에 연관 관계가 성립되어 있지 않다면 기존의 프로브의 정보만으로 최신의 정보를 얻기 힘들다.

<11> 올리고뉴클레오티드 프로브를 선정하는 방법은 마이크로어레이 뿐만 아니라 혼성화나 PCR 프리머 설계 등에서도 비슷하게 응용될 수 있는 분야로서 많은 연구가 이루어지고 있는 분야이다.

<12> 호주공개특허번호 제AU7534901호에는 열역학적 인자들(parameters)에 관련된 데이터베이스를 가지고, 서열정보, 혼성화 상태정보, 및 보정정보를 입력하였을 때 혼성화 열 특성을 예측해 내는 기술을 개시하고 있다. 즉, 호주공개특허번호 제AU7534901호에 개시된 발명은 핵산 혼성화 작업을 정확하게 예측할 수 있는 방법에 관한 것으로 정확한 프로브 설계 시스템의 구축에 이용될 수 있으나, 설계된 프로브를 이용하여 데이터를 검색하고 관련정보를 효율적으로 관리하는 것과는 무관하다.

<13> 한편, 유럽공개특허번호 제EP1103910호에는 유전자형 분류를 하고자 하는 영역을 미리 알고 있는 템플릿 서열에서 올리고뉴클레오티드 혼성화 프로브 자동 선택 방법을 개시하고 있다. 특히, 이 발명은 DNA 상의 변이(mutation)를 확인하기 위한 프로브 설계 방법에 관한 것으로, 원형서열(wild-type sequence)과 변형서열(mutant-type sequence)이 있을 때 그 둘의 차이(mutation site) 부분을 정의하고, 차이 부분을 중심으로 양 방

향으로 길이를 늘여가면서 두 서열을 구분할 수 있는 정도의 혼성화 특성 차이를 갖는 올리고뉴클레오티드를 선택하도록 하는 방법을 개시하고 있다. 이러한 방법은 변이위치(mutation site)를 확인하기 위한 올리고뉴클레오티드를 설계할 때 최적의 프로브들을 찾기 위해 적용될 수 있으나, 설계한 프로브의 관리, 변형서열과 원형서열의 관리에 대해서 제시하고 있지 않다.

<14> 미국특허번호 제US6403314호에는 단편화된 혼성화 예측 및 잠정적인 상호 교배 확인을 위한 계산방법 및 시스템이 개시되어 있다. 미국특허번호 제US6403314호에 개시된 발명은 프로브 집합과 샘플 집합을 구성하고 이들 사이에 가능한 결합 가능성을 모두 고려하여 원하는 샘플과 프로브가 결합하지 않는 교차 혼성화를 예측하는 방법에 관한 것이다. 프로브가 원하는 대상과 결합하지 않는 교차 혼성화의 문제는 프로브의 성능을 결정하는 주요 요소 가운데 하나로서 프로브를 설계하는 과정에서 반드시 검토해야 할 부분이다. 그러나, 샘플 집합과 프로브 집합, 그리고 샘플 집합과 샘플 집합 사이의 관계를 효과적으로 관리하지 못하면 샘플 집합이나 프로브 집합에에 약간의 변화만 있어도 이전에 구현한 정보가 모두 쓸모없게 된다.

<15> 미국특허번호 제US6251588호에는 올리고뉴클레오티드 프로브 서열 추정방법이 개시되어 있다. 미국특허번호 제US6251588호에 개시된 발명은 클러스터링(clustering) 기법을 통하여 혼성화 가능성을 예측하고, 프로브들의 순위를 결정하는 방식에 대한 것이다. 앞에서 언급한 방법들이 주로 열역학적 특성을 정확히 예측하고자 하는 데 초점을 맞추고 있었다면 이 방법은 보다 거시적인 관점에서 프로브들의 특성을 분석하고 효과적인 프로브들의 집합을 선정하기 위한 방법이라고 할 수 있다. 하지만 이 방법도 위에서 언급한 특허들과 같이 정보의 효과적인 관리에 대해서는 언급하고 있지 않다.

<16> 정리하면, 프로브 설계에 관련된 선행기술들은 주로 프로브 자체의 특성을 얼마나 더 정확하게 예측하고, 이를 바탕으로 더 양호한 프로브들을 선별할 것인가 하는 문제에만 관심이 있다고 할 수 있다. 한 번만 설계를 하고 다시는 관련 정보를 찾아보지 않아도 되는 상황과 달리, 관련 정보들이 계속 변화하고 새로운 정보들이 계속 알려지는 상황에서는 프로브들을 정확히 설계하는 방법뿐만 아니라 설계된 프로브 정보를 어떻게 하면 효과적으로 관리하여 이후에도 관련 정보를 쉽게 찾아볼 수 있는지 하는 문제가 될 수 있다.

<17> 미국특허번호 제6188783호에는 프로브 어레이칩 데이터베이스 제공방법 및 장치가 개시되어 있다. 미국특허번호 제6188783호에 개시된 발명은 데이터 관리에 초점을 두고 있는 발명이라고 할 수 있는데, 프로브와 샘플의 관계를 관계데이터베이스로 관리하는 것을 그 내용으로 하고 있다. 그러나, 하나의 프로브 정보가 여러 개의 샘플과 관계를 맺고 있을 때 샘플과 샘플 사이의 관계를 정의함으로써 프로브와 샘플의 새로운 관계를 정립할 수 있는 기능은 언급되어 있지 않다. 즉, 프로브와 샘플의 정보를 관계데이터베이스에 관리하더라도 샘플과 샘플 사이의 관계가 정의되어 있지 않으면 이전에 설계된 프로브를 가지고 최신 샘플과의 관계를 찾기가 쉽지 않다.

<18> 생물정보학(Bioinformatics)에서는 다양한 형태의 많은 데이터들을 다루기 때문에 데이터통합이 상당히 중요하게 여겨지고 있으며, 이와 관련된 특허들도 상당수 존재한다. 본 발명의 특성이 프로브와 샘플 서열 정보를 어떻게 효과적으로 관리하여 필요한 정보를 쉽게 얻을 것인가에 관계된 것이기 때문에 데이터통합과 관련된 기술도 검토를 해볼 필요가 있다.

<19> 우선, 국제공개번호 제WO0155911호에는 생물의학 자원에의 통합접근 시스템이 개시되어 있다. 국제공개번호 제WO0155911호에 개시된 발명은 로컬에 존재하는 데이터베이스와 원격에 존재하는 데이터베이스의 통신을 위하여 데이터 객체 연결자, 데이터 객체 결정자, 및 데이터 객체에 대한 GUI를 사용하는 데이터 가공 시스템에 관한 것이다.

<20> 다음으로, 국제공개번호 제WO0239468호에는 생물학 정보에 대한 통합시스템이 개시되어 있다. 국제공개번호 제WO0239468호에 개시된 발명은 UI를 갖는 클라이언트 환경과 소프트웨어 컴포넌트 사이의 통신을 담당하는 클라이언트 버스를 이용하여 인터페이스 기반의 객체 데이터 모델을 통합하는 방법에 관한 것이다.

<21> 다음으로, 국제공개번호 제WO0101294호에는 생물학 데이터 처리방법이 개시되어 있다. 국제공개번호 제WO0101294호에 개시된 발명은 여러 개의 데이터베이스 또는 서버에 대해 질의(query)를 할 때 구조화된 형식(예를 들면, XML)으로 입력하고, 이를 각각의 서버에서 요구하는 질의 형식으로 번역해 주는 번역서버로 데이터를 통합하는 방법에 관한 것이다.

<22> 마지막으로, 유럽공개특허번호 제EP1215614호에는 유전인자 분석 데이터 기록방법이 개시되어 있다. 유럽공개특허번호 제EP1215614호에 개시된 발명은 서열화 등의 실험적인 분석 방법을 통해 알아낸 유전자 변이 정보를 참조(reference) 정보와 함께 저장하는 방법에 관한 것이다. 그러나, 데이터의 항목 및 저장방법을 나열하고 있을 뿐 데이터를 특정한 구조로 저장하고 관리하는 방법에 대해서는 제시되어 있지 않다.

<23> 이들 특허의 내용은 주로 통합기술 자체에 있으며 검색 등 특정 업무를 수행하는 순간에 데이터들을 효과적으로 연결하여 보여줄 것인지에 관한 것이라 할 수 있다. 즉,

이러한 기술에도 이전의 정보를 추적하고 이전 정보와 최신 정보를 연결해 주는 방안에 대해서는 언급하고 있지 않다.

【발명이 이루고자 하는 기술적 과제】

<24> 본 발명이 이루고자 하는 기술적 과제는, 마이크로 어레이의 실험 결과를 이용하는 연구 및 진단에 있어서 올리고뉴클레이티드 프로브를 설계하기 위해 필요한 정보들을 통합하고 이전에 설계된 프로브 정보를 기초로 새로운 프로브 어레이를 설계할 수 있는 시스템 및 방법을 제공하는 데 있다.

<25> 본 발명이 이루고자 하는 다른 기술적 과제는, 마이크로 어레이의 실험 결과를 이용하는 연구 및 진단에 있어서 올리고뉴클레이티드 프로브를 설계하기 위해 필요한 정보들을 통합하고 이전에 설계된 프로브 정보를 기초로 새로운 프로브 어레이를 설계할 수 있는 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체를 제공하는 데 있다.

【발명의 구성 및 작용】

<26> 상기의 기술적 과제를 달성하기 위한, 본 발명에 따른 프로브 어레이 설계 시스템은, 게놈서열의 버전별 갱신이력이 기록된 크로스링크 맵이 저장되는 저장부; 상기 크로스링크 맵으로부터 상기 게놈서열을 구성하는 유전정보중에서 대상유전정보의 식별자와 관련있는 유전정보의 식별자 및 서열정보를 획득하는 정보검색부; 및 개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하고 상기 크로스링크 맵을 기초로 상기 기준유전정보의 시작 및 종료위치의 차이값를 계산하여 상기 기준유전

정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 위치추정부;를 구비한다.

<27> 상기의 다른 기술적 과제를 달성하기 위한, 본 발명에 따른 프로브 어레이 설계 방법은, 게놈서열의 버전별 갱신이력이 기록된 크로스링크 맵을 작성하는 단계; 상기 크로스링크 맵으로부터 상기 게놈서열을 구성하는 유전정보중에서 대상유전정보의 식별자와 관련있는 유전정보의 식별자 및 서열정보를 획득하는 단계; 개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하는 단계; 상기 크로스링크 맵을 기초로 상기 기준유전정보의 시작 및 종료위치의 차이값를 계산하는 단계; 및 상기 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 단계;를 포함한다.

<28> 이에 의해, 최근에 설계된 프로브에 대해서 최신의 정보를 제공할 수 있으며, 프로브를 설계할 당시의 게놈서열정보 및 식별자정보가 현재 시점에서 최신의 정보가 아니더라도 크로스링크 맵을 이용하여 최신의 정보를 찾아낼 수 있다.

<29> 이하에서 첨부된 도면들을 참조하여 본 발명에 따른 프로브 어레이 설계 시스템 및 방법의 바람직한 실시예에 대해 상세하게 설명한다.

<30> 도 1은 본 발명에 따른 프로브 어레이 설계 시스템에 대한 일 실시예의 구성을 도시한 블록도이다.

<31> 도 1을 참조하면, 본 발명에 따른 프로브 어레이 설계 시스템(100)은 저장부(110), 정보검색부(120), 위치추정부(130), 출력부(140) 및 정보통합부(150)로 구성된다.

<32> 저장부(110)에는 상이한 소스 사이의 관계를 정의하는 크로스링크 맵이 저장된다. 크로스링크 맵은 여러가지 식별자 정보를 이용하여 소스 사이의 관계를 정의하는 테이블의 형태로 작성된다. 예를 들어, 게놈서열 버전1.0과 게놈서열 버전2.0이 있을 때, 크로스링크 맵에는 각각의 단백질, 코딩서열(coding sequence : CDS), 엑손(exon)/인트론(intron), 조절영역(regulatory region), mRNA, 서열표지부위(sequence tagged site : STS), 발현유전자단편(expressed sequence tag : EST), 콘티그서열(contig sequence) 등의 식별자 위치정보를 이용하여 상호 참조를 할 수 있는 정보가 기록된다.

<33> 여기서, 위치정보는 각각의 식별자 정보를 기준으로 고려할 수 있는데, 예를 들어 특정한 프로브가 집합된 염색체정보 상에서는 어느 위치에 있는지, 콘티그정보 상에서는 어떤 위치에 있는지, 관련된 mRNA에서는 어떤 위치에 있는지, 주변에 알려져 있는 STS, EST, exon/intron 등을 봤을 때 상대적으로 어떤 위치에 있는지 등과 관련된 정보이다. 다양한 식별자정보를 기준으로 위치를 고려한다는 것은 다른 게놈서열(이는 새로운 정보를 바탕으로 같은 방식으로 새롭게 구성된 것일 수도 있고, 완전히 다른 방식으로 데이터를 구성한 것일 수도 있다. NCBI에서 발표한 build 28 정보와 build 29 정보의 차이는 전자에 해당하며, NCBI와 UCSC에서 발표한 정보의 차이는 후자에 해당한다) 사이에 정보를 비교할 때 가장 핵심적인 부분이라고 할 수 있다. 도 2에는 크로스링크 맵의 일 예가 도시되어 있다.

<34> 정보검색부(120)는 크로스링크 맵에서 식별자정보 사이의 관계를 바탕으로 정보를 추적한다. 정보검색부(120)에서 수행하는 기능 중의 하나는 특정 정보에 대해 요청이 들어왔을 때 그와 관련된 정보를 모두 검색하여 출력하는 기능이다. 예를 들면, BRCA2 유전자의 mRNA에 대해 요청이 들어오면, 정보검색부(120)는 해당 정보를 담고 있는 크로스

링크 맵상의 모든 기록을 검색하여 추출한다. 만약, 요청받은 mRNA 정보가 특정 게놈집합 버전에 대한 것이었다면 정보검색부(120)는 크로스링크 맵에 있는 정보들을 확인하여 그 정보가 최신 정보인지 확인하고, 최신 정보가 아니라면 정보가 갱신되었다는 점도 같이 출력한다.

<35> 위치추정부(130)는 동일한 정보에 대해 여러 가지 결과가 나왔을 때 그 가운데 우선순위를 정하여 정확한 결과를 예측한다. 크로스링크 맵에는 특정한 유전정보에 대해 다양한 기준으로 위치정보가 기록되어 있으므로, 위치추정부(130)는 동일한 정보에 대해 우선 순위를 정하여 정확히 예측할 수 있다. 즉, 위의 예에서처럼 BRCA2 유전자의 mRNA에 대해 요청이 들어오면 해당 정보의 위치를 염색체, 관련 콘티그, 코딩서열, 단백질서열, exon/intron 정보 등과 같은 다양한 기준에 의해 크로스링크 맵을 검색하게 된다.

<36> 이 때, 위치추정부(130)는 개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하고 크로스링크 맵을 기초로 기준유전정보의 시작 및 종료위치의 차이값를 계산하여 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 대상유전정보의 위치로 결정한다. 일반적으로 엑손이나 단백질은 특정 개체의 게놈에 모두 포함되어 있으며, 이러한 유전정보들은 게놈서열의 버전이 갱신되어도 위치의 변화가 크지 않다. 위치추정부(130)는 버전의 갱신시에도 위치변화가 적은 유전정보들을 기준유전정보로 설정하고 설정된 기준유전정보 중에서 개체의 보유율이 높은 유전정보에 대해 계산된 차이값에 우선순위를 부여하여 대상유전정보의 위치를 결정한다.

<37> 한편, 위치추정부(130)는 대상유전정보가 존재할 가능성이 있는 영역을 설정한 후 해당 영역내에서 대상유전정보의 위치를 결정할 수도 있다. 이 경우, 위치추정부(130)는

추정영역설정부(132) 및 위치결정부(134)를 구비한다. 위치추정부(130)의 상세한 구성은 도 3에 도시되어 있다. 추정영역설정부(132)는 크로스링크 맵을 기초로 기준그룹에서 제외된 유전정보의 시작 및 종료위치의 차이값를 계산하고 기준그룹에서 제외된 유전정보에 대해 계산된 차이값을 기초로 게놈서열상에서의 대상유전정보의 위치에 대한 추정영역을 설정한다. 위치결정부(134)는 추정영역내에서 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 대상유전정보의 위치로 결정한다.

<38> 또한, 위치추정부(130)는 크로스링크 맵을 기초로 기준그룹을 갱신하는 갱신부(136)를 구비한다. 갱신부(136)는 게놈서열에 대한 각각의 버전에 공통으로 존재하는 유전정보의 시작위치 및 종료위치의 차이값을 계산한 후 계산된 차이값이 소정 범위내에 존재하는 유전정보를 선정하여 기준그룹을 갱신한다. 게놈서열에 대한 각각의 버전에 공통으로 존재하는 유전정보의 시작위치 및 종료위치의 차이값은 갱신부(136)에서 계산될 수 있으며, 이와 달리, 추정영역설정부(132)에서 계산되어 갱신부(136)로 입력될 수도 있다.

<39> 출력부(140)는 요청받은 정보와 최신 정보와의 차이를 출력한다. 크로스링크 앱의 검색시 상이한 게놈버전에 대한 데이터를 요청받은 경우 위치정보가 일치하지 않을 수 있다. 예를 들어, 염색체상에서는 10,000bp(base pair)가 이동하였는데 콘티그상에서는 9,900bp만 이동할 수 있다. 일반적으로, 기능과 관련된 단백질서열 및 exon 정보가 염색체나 콘티그정보에 비해 위치정보가 잘 보존되므로, 위치가 잘 보존되는 정보에 가중치를 두어 상이한 게놈버전 사이에서도 정확한 위치를 찾을 수 있다.

<40> 정보통합부(150)는 다양한 정보들에서 크로스링크 맵에 필요한 정보를 취한다. 다양한 소스들로부터 크로스링크 맵에 필요한 정보를 얻기 위해서는 각각의 데이터 형식을

크로스링크 맵의 데이터 형식으로 변환해 주는 구성요소가 필요하다. 정보통합부(150)는 새로운 데이터를 고려하게 될 때마다 기존의 데이터를 새로운 데이터의 형식에 맞게 변환한다. 데이터가 분산되어 있는 경우에도 정보통합부(130)는 해당 데이터 제공자에 접근하여 원하는 형태로 데이터를 가져온다.

<41> 도 4는 본 발명에 따른 프로브 설계 방법의 수행과정을 도시한 흐름도이다.

<42> 도 4를 참조하면, 먼저 프로브를 설계할 대상유전정보를 외부로부터 입력받는다(S400). 이 때, 프로브의 설계목적이 염기서열의 변이(mutation)를 확인하기 위한 것인 경우에는 유전자와 함께 관련된 변이정보가 입력된다.

<43> 정보검색부(120)는 크로스링크 맵에서 입력된 대상유전정보에 대응하는 게놈서열 및 식별자정보를 검색한다(S410). 다음으로, 정보검색부(120)는 크로스링크 맵에서 확인된 식별자정보를 기초로 대상유전정보와 관련있는 변이정보를 확인한다(S420). 그리고 나서, 정보검색부(120)는 확인된 변이정보 및 식별자정보가 최신 정보들인지 확인한다(S430). 크로스링크 맵은 항상 최신 정보를 포함하고 있으므로, 크로스링크 맵에 기록되어 있는 정보만을 가지고 확인된 정보들이 최신 정보인가를 용이하게 파악할 수 있다.

<44> 변이 및 식별자정보가 최신 정보로 확인되면, 일반적인 프로브 설계 방법에 의해 변이정보를 게놈서열에 위치시키고 프로브 설계에 필요한 기타 다른 인자들을 전달받아 프로브를 설계한다(S440). 그러나, 변이 및 식별자정보가 최신 서열에 관한 정보가 아니라면 바로 프로브 설계에 들어갈 수 없다. 이 경우, 위치추정부(130)는 크로스링크 맵을 이용하여 해당되는 변이 및 식별자정보를 기초로 최신 서열상에서의 대상유전정보의 위치를 추정한다(S450). 이러한 추정과정을 수행한 이후에 최신 서열과 그 위에 표시된 변이 및 식별자정보에 의해 프로브를 설계한다(S460).

<45> 도 5는 본 발명에 따른 프로브 설계 방법의 위치추정과정을 도시한 흐름도이다.

<46> 도 5를 참조하면, 정보검색부(120)는 정보를 얻고자 하는 프로브를 선택하고, 선택한 프로브가 어떠한 게놈서열을 기반으로 설계된 것인지 파악한다(S500). 파악된 정보는 프로브를 설계할 때 관련 정보로 관리된다. 그리고, 현재 사용 가능한 서열이 최신 정보인가를 확인한다(S510).

<47> 정보검색부(120)는 프로브가 현재 사용가능한 최신의 서열로 설계되어 있다면 크로스링크 맵에서 최신 서열과 관련된 식별자정보들을 취하여 해당 정보들을 출력한다(S520). 이 때, 이전 서열 또는 단백직서열과 같은 다른 형태의 정보 역시 크로스링크 맵의 정보를 이용하여 최신 서열정보와 함께 출력한다.

<48> 이와 달리, 프로브가 현재 사용가능한 최신의 서열이 아닌 이전의 서열로 설계된 프로브로 확인되면, 위치추정부(130)는 크로스링크 맵을 이용하여 프로브가 최신 서열의 어느 부분에 위치하는지를 추정하는 과정을 수행한다. 먼저, 위치추정부(130)는 정보검색부(120)에서 검색된 이전 위치를 기준으로 해당 위치를 포함하는 최신 서열의 식별자정보들을 입력받는다(S530). 정보검색부(120)에서 검색된 식별자정보에는 이전에 프로브가 속했던 콘티그, 엑손, 단백질서열, mRNA 등이 포함된다. 도 6에는 대상유전정보가 BRCA2일 때, 크로스링크 맵에서 얻어진 이전 서열과 최신 서열의 식별자정보의 시작 및 종료위치가 도시되어 있다.

<49> 다음으로, 위치추정부(130)는 입력된 식별자정보들을 최신 서열의 식별자정보로부터 찾아내어 이들 사이의 관계를 파악한다(S540). 이를 위해, 위치추정부(130)는 이전 서열과 최신서열상에서 대응되는 식별자정보의 시작 및 종료위치의 차이값을 계산한다.

도 6을 참조하면, SNPN을 제외한 나머지 유전정보의 차이값은 -85668~-85669의 범위에 있음을 알 수 있다.

<50> 이 때, 여러 개의 식별자정보가 서로 충돌할 수 있다. 즉, 콘티그상의 위치 차이와 엑손사이의 위치 차이가 다를 수 있다. 이 때, 위치추정부(130)는 보다 신빙성이 높은 정보를 우선으로 하여 정보들을 종합하고 가장 정확할 것으로 예측되는 위치 변경 정보를 이용하여 최신 서열상에서의 대상유전정보의 위치를 추정한다(S550). 이 때, 일반적으로 개체의 모든 게놈서열에 포함되어 있는 엑손을 기준으로 위치를 추정한다. 이 경우, UCSC.200104버전에 존재하는 엑손은 -85668~-85669의 범위에서 이동하였으므로, UCSC.200104버전에 존재하던 SNPN은 UCSC.200206버전에 존재하지 않는 것으로 추정할 수 있다. 만약, 위치를 예측하기 힘들 정도로 식별자정보들 사이의 차이가 심하면 그 결과를 출력하고 매핑이 불가능함을 출력할 수도 있다.

<51> S550단계에서 위치추정부(130)는 이전 서열과 최신서열상에서 대응되는 식별자정보의 시작 및 종료위치의 차이값을 기초로 잠정적인 추정영역을 설정할 수도 있다. 이 경우, CDS의 차이는 -148990~-149048이므로 위치변화가 가장 큰 유전정보이다. 따라서, 위치추정부(130)는 각각의 유전정보중에 대해 계산된 차이값이 존재하는 범위내에서 대상유전정보의 위치를 추정하게 된다.

<52> 상술한 방식으로 프로브의 위치를 최신 서열에 매핑하면 해당 영역에 관련된 최신의 식별자정보를 얻을 수 있다. 그리고, 이들 정보를 종합적으로 이용하면 해당 프로브에 관한 다양한 최신 정보를 얻을 수 있게 된다. 이러한 기능에 의해 프로브 정보의 사용자가 프로브를 새롭게 설계를 하지 않고도 새롭게 설계를 한 프로브로부터 얻을 수 있는 것에 해당하는 정보를 얻을 수 있다. 예를 들어, 3년 전의

마이크로 어레이정보와 최근의 마이크로 어레이정보를 비교해 보고자 할 때 각각의 프로브가 달라질 수도 있지만, 동일한 프로브라도 관련 정보가 현저히 차이날 수 있다. 종래에는 프로브의 위치를 최신 정보에서 별도로 검색하여 이러한 정보를 확인하였다. 물론, 비슷한 영역의 정보는 유전자명 등의 관련 정보를 이용하여 찾을 수 있으나, 정확한 위치 정보를 찾기 위해서는 추가적인 노력이 요구된다. 그러나, 본 발명에 따른 프로브 설계 시스템에서는 이전 정보와 최신 정보 사이에 정의된 각종 식별자 정보들을 이용하여 그 위치를 추가적인 노력 없이 정확히 예측해 낼 수 있다.

<53> 또한, 하나의 대상(예를 들면 특정 유전자)에 대해 여러 가지 종류의 마이크로 어레이 실험 결과가 존재하는 경우 본 발명을 활용하면 프로브의 정보를 추적할 수 있으므로, 이전 실험 정보와 최신 실험 정보를 비교할 수 있고 이전 실험 정보에 대해 최신의 관련 정보를 찾아볼 수 있다. 예를 들어, 질병을 진단하기 위한 프로브를 설계하고자 할 때, 질병에 관계된 변이 및 유전자정보를 관리하는 별도의 데이터베이스에서 획득된 관련 정보를 크로스링크 맵을 기초로 비교하여 게놈서열상의 관련 정보를 얻을 수 있다. 이러한 기능은 프로브 설계를 위한 관련 정보를 보다 쉽게 관리할 수 있다는 장점을 갖게 한다. 또한 이전 정보를 기준으로 명명된 정보라도 최신 정보와 비교해 볼 수 있다는 장점도 가진다.

<54> 본 발명은 또한 컴퓨터로 읽을 수 있는 기록매체에 컴퓨터가 읽을 수 있는 코드로서 구현하는 것이 가능하다. 컴퓨터가 읽을 수 있는 기록매체는 컴퓨터 시스템에 의하여 읽혀질 수 있는 데이터가 저장되는 모든 종류의 기록장치를 포함한다.

컴퓨터가 읽을 수 있는 기록매체의 예로는 ROM, RAM, CD-ROM, 자기 테이프, 플로피디스크, 광데이터 저장장치 등이 있으며, 또한 캐리어 웨이브(예를 들어 인터넷을 통한 전송)의 형태로 구현되는 것도 포함한다. 또한 컴퓨터가 읽을 수 있는 기록매체는 네트워크로 연결된 컴퓨터 시스템에 분산되어 분산방식으로 컴퓨터가 읽을 수 있는 코드가 저장되고 실행될 수 있다.

<55> 이상에서 본 발명의 바람직한 실시예에 대해 도시하고 설명하였으나, 본 발명은 상술한 특정의 바람직한 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변형 실시가 가능한 것은 물론이고, 그와 같은 변경은 청구범위 기재의 범위 내에 있게 된다.

【발명의 효과】

<56> 본 발명에 따른 프로브 설계 시스템 및 방법에 의하면, 최근에 설계된 프로브에 대해서 최신의 정보를 제공할 수 있으며, 프로브를 설계할 당시의 게놈서열정보 및 식별자정보가 현재 시점에서 최신의 정보가 아니더라도 크로스링크 맵을 이용하여 최신의 정보를 찾아낼 수 있다. 또한, 본 발명은 마이크로 어레이의 성능 향상을 위해 지속적으로 프로브 정보가 변경되는 경우에 적용될 수 있으며, 게놈서열정보나 관련 식별자정보가 프로브 설계 수단과 분리되어 있기 때문에 외부의 잘 관리된 데이터를 그대로 쓸 수 있다는 장점이 있다.

【특허청구범위】

【청구항 1】

게놈서열의 버전별 갱신이력이 기록된 크로스링크 맵이 저장되는 저장부;

상기 크로스링크 맵으로부터 상기 게놈서열을 구성하는 유전정보중에서 대상유전정보의 식별자와 관련있는 유전정보의 식별자 및 서열정보를 획득하는 정보검색부; 및

개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하고 상기 크로스링크 맵을 기초로 상기 기준유전정보의 시작 및 종료위치의 차이값를 계산하여 상기 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 위치추정부;를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 시스템.

【청구항 2】

제 1항에 있어서,

상기 게놈서열에 대한 다양한 소스들로부터 상기 크로스링크 맵에 기록되는 엔트리에 대응되는 데이터 상기 크로스링크 맵의 기록형식으로 변환하는 정보통합부를 더 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 시스템.

【청구항 3】

제 1항에 있어서,

상기 크로스링크 맵에 기록되는 엔트리는 상기 게놈서열의 명칭, 상기 게놈서열의 버전, 상기 게놈서열을 구성하는 유전정보의 식별자, 상기 게놈서열을 구성하는 유전정보의 상기 게놈서열상에서의 시작위치와 종료위치, 및 상기 게놈서열을 구성하는 유전정

보의 길이를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 시스템.

【청구항 4】

제 1항에 있어서,

상기 위치추정부는 상기 기준유전정보 중에서 개체의 보유율이 높은 유전정보에 대해 계산된 차이값에 우선순위를 부여하여 상기 대상유전정보의 위치를 결정하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 시스템.

【청구항 5】

제 1항에 있어서,

상기 위치추정부는,

상기 크로스링크 맵을 기초로 상기 기준그룹에서 제외된 유전정보의 시작 및 종료 위치의 차이값를 계산하고 상기 기준그룹에서 제외된 유전정보에 대해 계산된 차이값을 기초로 상기 게놈서열상에서의 상기 대상유전정보의 위치에 대한 추정영역을 설정하는 추정영역설정부; 및

상기 추정영역내에서 상기 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 위치결정부;를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 시스템.

【청구항 6】

제 1항에 있어서,

상기 위치추정부는 상기 게놈서열에 대한 각각의 버전에 공통으로 존재하는 유전정보의 시작위치 및 종료위치의 차이값을 계산한 후 계산된 차이값이 소정 범위내에 존재하는 유전정보를 선정하여 상기 기준그룹을 갱신하는 갱신부를 더 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 시스템.

【청구항 7】

게놈서열의 버전별 갱신이력이 기록된 크로스링크 맵을 작성하는 단계;

상기 크로스링크 맵으로부터 상기 게놈서열을 구성하는 유전정보중에서 대상유전정보의 식별자와 관련있는 유전정보의 식별자 및 서열정보를 획득하는 단계;

개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하는 단계;

상기 크로스링크 맵을 기초로 상기 기준유전정보의 시작 및 종료위치의 차이값를 계산하는 단계; 및

상기 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 단계;를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 방법.

【청구항 8】

제 7항에 있어서,

상기 크로스링크 맵에 기록되는 엔트리는 상기 게놈서열의 명칭, 상기 게놈서열의 버전, 상기 게놈서열을 구성하는 유전정보의 식별자, 상기 게놈서열을 구성하는 유전정보의 상기 게놈서열상에서의 시작위치와 종료위치, 및 상기 게놈서열을 구성하는 유전정

보의 길이를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 방법.

【청구항 9】

제 7항에 있어서,

상기 위치결정단계는 상기 기준유전정보 중에서 개체의 보유율이 높은 유전정보에 대해 계산된 차이값에 우선순위를 부여하여 상기 대상유전정보의 위치를 결정하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 방법.

【청구항 10】

제 7항에 있어서,

상기 게놈서열에 대한 각각의 버전에 공통으로 존재하는 유전정보의 시작위치 및 종료위치의 차이값을 계산한 후 계산된 차이값이 소정 범위내에 존재하는 유전정보를 선정하여 상기 기준그룹을 갱신하는 단계;를 더 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 방법.

【청구항 11】

제 7항에 있어서,

상기 위치결정단계는,

상기 크로스링크 맵을 기초로 상기 기준그룹에서 제외된 유전정보의 시작 및 종료위치의 차이값를 계산하고 상기 기준그룹에서 제외된 유전정보에 대해 계산된 차이값을 기초로 상기 게놈서열상에서의 상기 대상유전정보의 위치에 대한 추정영역을 설정하는 단계; 및

상기 추정영역내에서 상기 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 단계;를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 방법.

【청구항 12】

게놈서열의 버전별 갱신이력이 기록된 크로스링크 맵을 작성하는 단계;

상기 크로스링크 맵으로부터 상기 게놈서열을 구성하는 유전정보중에서 대상유전정보의 식별자와 관련있는 유전정보의 식별자 및 서열정보를 획득하는 단계;

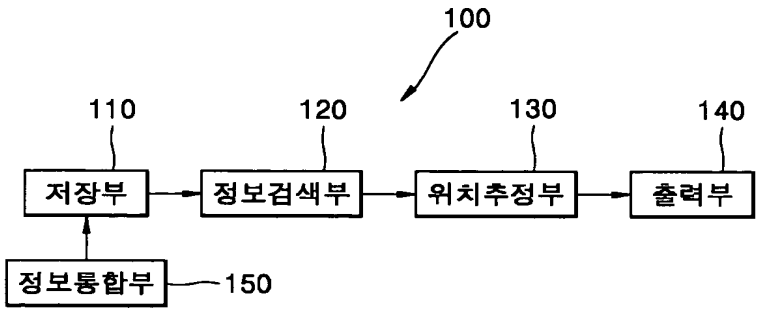
개체의 보유율이 소정의 기준값 이상인 기준유전정보로 선택하여 기준그룹을 결정하는 단계;

상기 크로스링크 맵을 기초로 상기 기준유전정보의 시작 및 종료위치의 차이값를 계산하는 단계; 및

상기 기준유전정보에 대해 계산된 차이값에 대응하는 위치를 게놈서열상에서의 상기 대상유전정보의 위치로 결정하는 단계;를 포함하는 것을 특징으로 하는 이형 유전정보를 이용한 프로브 어레이 설계 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터로 읽을 수 있는 기록매체.

【도면】

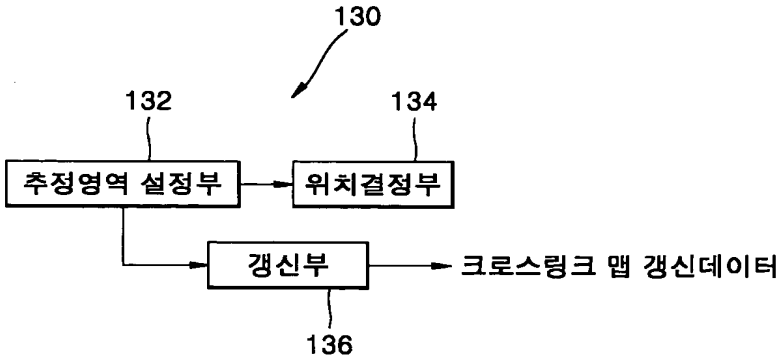
【도 1】



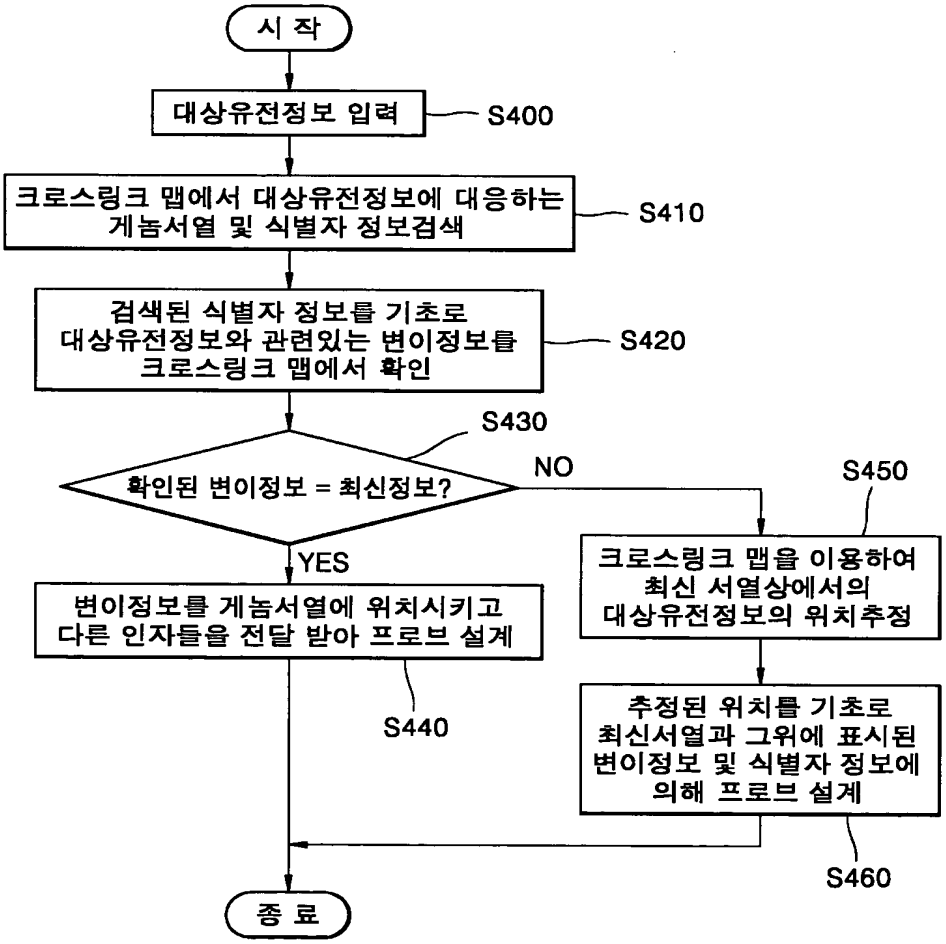
【도 2】

Genome ID	Genome Version	Chr	Chr Start	Chr End	Annot. Type	Annot. ID	Annot. Start	Annot. End	Annot. Orient
UCSC	200104	Chr13	30961644	31045834	TRANS	BRCA2.TRANS	1	84190	+
UCSC	200104	Chr13	30962625	31044936	CDS	BRCA2. CDS	1	82311	+
UCSC	200104	Chr13	30961644	30961832	EXON	BRCA2.EXON1	1	188	+
UCSC	200104	Chr13	30961820	30961821	SNP	dbSNP :206118	1	2	+
UCSC	200206	Chr13	30875976	30960165	TRANS	BRCA2.TRANS	1	84189	+
UCSC	200206	Chr13	30876957	30959267	CDS	BRCA2. CDS	1	82310	+
UCSC	200206	Chr13	30875976	30876164	EXON	BRCA2.EXON1	1	188	+
UCSC	200206	Chr13	30876151	30876152	SNP	dbSNP :206118	1	2	+

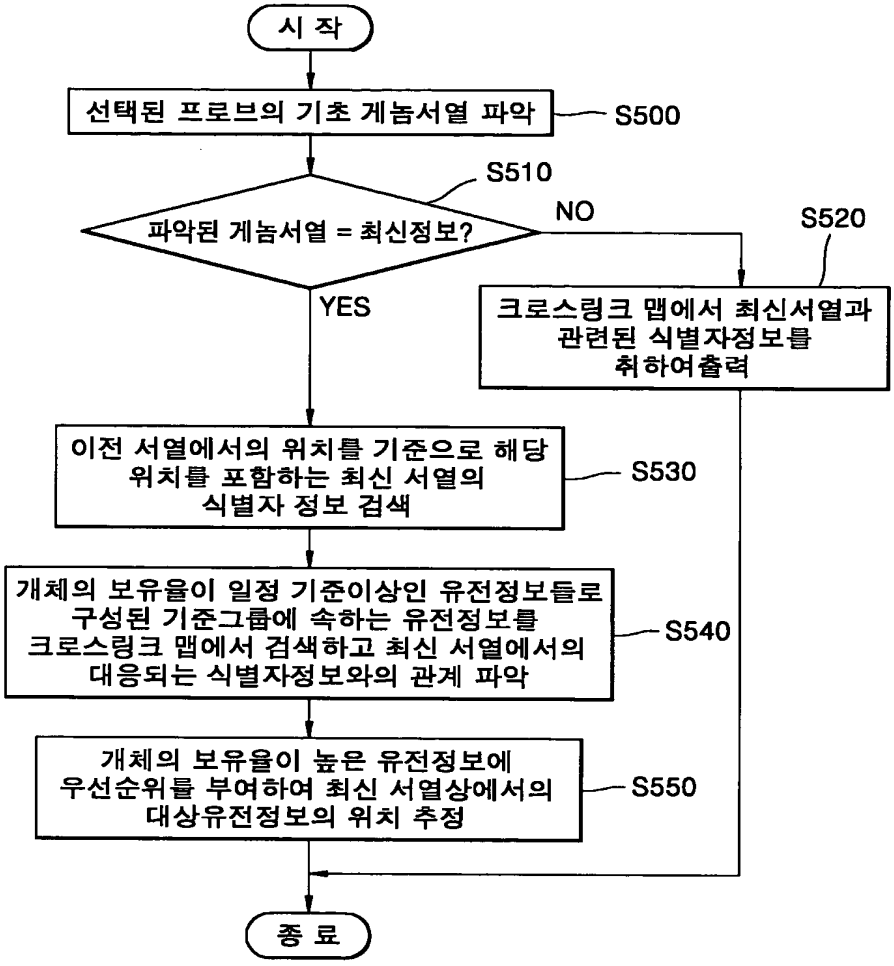
【도 3】


130
132
134
추정영역 설정부
위치결정부
갱신부
크로스링크 맵 갱신데이터
136

【도 4】


시 작
대상유전정보 입력
S400
크로스링크 맵에서 대상유전정보에 대응하는 게놈서열 및 식별자 정보검색
S410
검색된 식별자 정보를 기초로 대상유전정보와 관련있는 변이정보를 크로스링크 맵에서 확인
S420
S430
확인된 변이정보 = 최신정보?
NO
S450
YES
변이정보를 게놈서열에 위치시키고 다른 인자들을 전달 받아 프로브 설계
S440
크로스링크 맵을 이용하여 최신 서열상에서의 대상유전정보의 위치추정
추정된 위치를 기초로 최신서열과 그위에 표시된 변이정보 및 식별자 정보에 의해 프로브 설계
S460
종 료

【도 5】


시 작
선택된 프로브의 기초 게놈서열 파악
S500
S510
파악된 게놈서열 = 최신정보?
NO
S520
YES
크로스링크 맵에서 최신서열과 관련된 식별자정보를 취하여출력
이전 서열에서의 위치를 기준으로 해당 위치를 포함하는 최신 서열의 식별자 정보 검색
S530
개체의 보유율이 일정 기준이상인 유전정보들로 구성된 기준그룹에 속하는 유전정보를 크로스링크 맵에서 검색하고 최신 서열에서의 대응되는 식별자정보와의 관계 파악
S540
개체의 보유율이 높은 유전정보에 우선순위를 부여하여 최신 서열상에서의 대상유전정보의 위치 추정
S550
종 료

【도 6】

UCSC.200104
TRANS : 30961644 - 31045834 (84190)
CDS : 30962625 - 31044936 (82311)
EXON : 30961644 - 30961832 (188)
SNPN : 30961820 - 30961821 [206118]
EXON : 31025908 - 31026072 (164)
EXON : 31040851 - 31041096 (245)
SNPN : 31041006 - 31041007 [1799968]
EXON : 31043063 - 31043210 (147)
GCCTCATATGTTAATTGCTGCAAGCAACCTCCAGTGGCGAC
TAATTACTGCAAGCAACCTC
31041000 - 31041020
UCSC.200206
TRANS : 30875976 - 30960165 (84189)
CDS : 30876957 - 30959267 (82310)
EXON : 30875976 - 30876164 (188)
SNPN : 30876151 - 30876152 206118
EXON : 30876918 - 30877024 (106)
EXON : 30955185 - 30955430 (245)
EXON : 30957394 - 30957541 (147)
EXON : 30958658 - 30959707 (1049)
GCCTCATATGTTAATTGCTGCAAGCAACCTCCAGTGGCGAC
TAATTGCTGCAAGCAACCTC
30955325 - 30955345